The Nishimori line and Bayesian Statistics

Yukito Iba

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu Minato-ku,Tokyo 106-8569,Japan, iba@ism.ac.jp

Abstract. “Nishimori line” is a line or hypersurface in the parameter space of systems with quenched disorder, where simple expressions of the averages of physical quantities over the quenched random variables are obtained. It has been playing an important role in the theoretical studies of the random frustrated systems since its discovery around 1980. In this paper, a novel interpretation of the Nishimori line from the viewpoint of statistical information processing is presented. Our main aim is the reconstruction of the whole theory of the Nishimori line from the viewpoint of Bayesian statistics, or, almost equivalently, from the viewpoint of the theory of error-correcting codes. As a byproduct of our interpretation, counterparts of the Nishimori line in models without gauge invariance are given. We also discussed the issues on the “finite temperature decoding” of error-correcting codes in connection with our theme and clarify the role of gauge invariance in this topic.

Submitted to: *J. Phys. A: Math. Gen.*

1. Introduction

There are not many rigorous results that are useful for the study of random frustrated systems. Among them, theorems related to the Nishimori line of random spin models form an important family. There have been many papers [1, 2, 3, 4, 5] about the Nishimori line after the seminal paper [1] of Nishimori. There is, however, still a mystery about the Nishimori line, i.e., its physical meaning and the motivation behind the proof are not yet clear.

The purpose of this paper is to provide a novel interpretation on the Nishimori line from the viewpoint of statistical information processing, more specifically from Bayesian statistics or from the coding theory. Our interpretation has two advantages. First it gives an interesting example of an unexpected relation between two different areas, rigorous arguments in the statistical physics and Bayesian statistics, and elucidates the meaning of the trick in the derivation of the Nishimori line. Secondly it gives some new results on the analog of the Nishimori line without *gauge invariance* in the sense of Toulouse [6].

Our arguments are closely related to the works on the “the optimality of finite-temperature decoding” of error-correcting codes [7, 8, 9, 10] . In fact, some of our results are essentially given in Sourlas [10]. In these works, however, finite-temperature decoding is discussed with spin glass theory. On the other hand, our aim here is reverse the direction of the arguments and discuss the whole theory of the Nishimori line from the viewpoint of statistical science. We will discuss further on the finite-temperature decoding in the last section.

In this paper, we make efforts to give a self-contained description of this material. No special knowledge on Bayesian statistics, error-correcting codes and gauge invariance of spin glass is assumed.

2. Bayesian Framework

In this section, we give basic notions and terminology of Bayesian statistics. We also discuss identities and inequalities that naturally arise from the Bayesian framework. Although the motivation for these formulas as well as their proofs are quite simple, they are essential in the derivation of the properties of the Nishimori line.

Let us assume that our data y is generated by a probability distribution $p(y|x)$, which is parameterized by the value of an unknown variable x. In the Bayesian framework, we also assume that the parameter x is, in itself, a random sample from a *prior distribution* $\pi(x)$. With these assumptions, the probability distribution of the parameter x conditioned on given data y is

$$p(x|y) = \frac{p(y|x)\pi(x)}{\sum_x p(y|x)\pi(x)}. \tag{1}$$

Here $\sum_x$ means the summation or integral over the possible values of x. This distribution, the *posterior distribution*, is the source of knowledge with given data y in the Bayesian formalism.

Similar formalism is also used in seemingly different branch of the information science, the theory of error-correcting codes. Consider a noisy channel and a set of messages. We encode and send a message x through the noisy channel and someone at the other end of the channel tries to infer the original message x from the output y. If we assume that the probability $p(y|x)$ of an output y with the input x and the distribution $\pi(x)$ of the average frequencies of input messages, the conditional probability $p(x|y)$ of an input x with the output y is given by (1). Note that the probability $p(y|x)$ represents the coding scheme as well as the noise of the channel in this formalism.

We will introduce notations that indicate the averages over different types of distributions. Here the symbol $A(x)$ denotes a function of the parameter x and $B(y)$ denote a function of the data y . First we define the average over the prior distribution of x,

$$[A(x)]_{\pi(x)} = \sum_x A(x)\pi(x). \tag{2}$$

We also define the average over the posterior distribution of x,

$$\langle A(x)\rangle_{p(x|y)} = \sum_x A(x)p(x|y). \tag{3}$$

Finally we define the average over the probability distribution $p(y|x)$ of data y with the given parameter x,

$$[B(y)]_{p(y|x)} = \sum_y B(y)p(y|x). \tag{4}$$

These notations are not common in the literatures on Bayesian statistics. They are introduced to contrast the analogy to the statistical physics of systems with quenched disorder. Later we show that the averages [ ] correspond to the quenched average over the configuration of the impurities and $\langle\ \rangle$ corresponds to the thermal average.

Let us consider relations among these averages. First we note that an identity

$$[\langle A(x')\rangle_{p(x'|y)}]_{p(y|x)}]_{\pi(x)} = [A(x)]_{\pi(x)}, \tag{5}$$

holds. The posterior average $\langle A(x')\rangle_{p(x'|y)}$ can be regarded as an estimate of $A(x)$ from the data y. It is a random variable dependent on y and the identity (5) shows that the average of it over the possible values of the data y and the original parameter x coincides with the prior average $[A(x)]_{\pi(x)}$. The proof of the formula (5) is straightforward. When we substitute the left hand side of (5) for the definition of the averages (3), (4), (2), we obtain the expression,

$$[\langle A(x')\rangle_{p(x'|y)}]_{p(y|x)}]_{\pi(x)} = \sum_x \sum_y \frac{\sum_{x'} A(x')p(y|x')\pi(x')}{\sum_{x'} p(y|x')\pi(x')} \cdot p(y|x)\pi(x). \tag{6}$$

By changing the order of the summation and a dummy index, we can show that the factors $\sum_x p(y|x)\pi(x)$ in the numerator and denominator cancel each other. Using $\sum_y p(y|x') = 1$ and $\sum_{x'} A(x')\pi(x') = [A(x)]_{\pi(x)}$, the proof of (5) is completed.

It is easy to generalize (5) to an identity

$$[\langle C(x', y)\rangle_{p(x'|y)}]_{p(y|x)}]_{\pi(x)} = [[C(x, y)]_{p(y|x)}]_{\pi(x)}. \tag{7}$$

Here $C(x, y)$ is a function of the data (the output of the channel) y as well as the parameter x. The proof of the relation (7) is essentially the same as that of (5). The only difference from (5) is that the average $[\ ]_{p(y|x)}$ in the right hand side cannot be removed.

In these arguments, we assume that the "true" distributions $p(y|x)$ and $\pi(x)$ behind given data are exactly known. They are, however, often unknown in a real world example. A way to fill this gap is to include "hyper parameters" α and γ in the expression of $p(y|x)$ and $\pi(x)$ and estimate them from the data. Hereafter we use the notation $p_\alpha(y|x)$ and $\pi_\gamma(x)$ to indicate the distributions that contain hyperparameters. An approach to estimate hyperparameters α and γ from the data y is the minimization of a free-energy-like quantity,

$$F(\alpha, \gamma) = -\log \sum_x p_\alpha(y|x)\pi_\gamma(x). \tag{8}$$

Note that the procedure based on the *marginal likelihood* $\sum_x p_\alpha(y|x)\pi_\gamma(x)$ is successfully used by the many authors in practical problems. It is called by a lot of different terms, say, the maximization of *type II likelihood* [11, 12], the minimization of *ABIC* [13], the maximization of *evidence* [14, 15], and, simply, the maximization of the likelihood of α and γ [16, 17, 18] ‡.

At the moment, we assume that the form of the distribution $\pi_\gamma(x)$ and $p_\alpha(y|x)$ is correctly known except the values of the hyper parameters. Even in this case, the hyperparameters (α, γ) that maximize (8) are random variables dependent on the data y and they fluctuate around the true values (α_0, γ_0) of (α, γ). However, (α, γ)

‡ When the expression of the probability $p(y|x)$ is considered as the function of x with given data y, it is called "likelihood of the parameter x". This terminology is preferred by the non-Bayesians who do not treat the parameter x as a random variable, but is also used by Bayesians. The mixture distribution $\sum_x p_\alpha(x|y)\pi_\gamma(x)$ at the right hand side of (8) can be regarded as the likelihood of the hyperparameters α and γ.

that minimize the average of $F(\alpha,\gamma)$ over the true distribution $\pi_{\gamma_0}(x)$ and $p_{\alpha_0}(y|x)$ coincides with the true value (α_0,γ_0), i.e, the inequality

$$[[F(\alpha_0,\gamma_0)]_{p_{\alpha_0}(y|x)}]_{\pi_{\gamma_0}(x)} \leq [[F(\alpha,\gamma)]_{p_{\alpha_0}(y|x)}]_{\pi_{\gamma_0}(x)} \tag{9}$$

holds for any value of α and γ.

If the right hand side of (9) is a sufficiently smooth function of (α,γ), the derivatives of F at $(\alpha,\gamma) = (\alpha_0,\gamma_0)$ should be zero. For example, the following relations are direct consequences of (9).

$$[[\,\frac{\partial}{\partial\alpha}F(\alpha,\gamma)\,]_{p_{\alpha_0}(y|x)}]_{\pi_{\gamma_0}(x)}\Big|_{(\alpha,\gamma)=(\alpha_0,\gamma_0)} = 0. \tag{10}$$

$$[[\,\frac{\partial}{\partial\gamma}F(\alpha,\gamma)\,]_{p_{\alpha_0}(y|x)}]_{\pi_{\gamma_0}(x)}\Big|_{(\alpha,\gamma)=(\alpha_0,\gamma_0)} = 0. \tag{11}$$

Here, the derivatives $\partial F/\partial\alpha$ and $\partial F/\partial\gamma$ should be interpreted as the derivatives $\partial F/\partial\alpha_k$ and $\partial F/\partial\gamma_k$ with each component of $\alpha = \{\alpha_k\}$ and $\gamma = \{\gamma_k\}$, when α and γ are vectors with more than one components. The conditions on the second derivatives are also derived from (9) by using positive semi-definiteness of the Hessian, say,

$$[[\,\frac{\partial^2}{\partial^2\alpha}F(\alpha,\gamma)\,]_{p_{\alpha_0}(y|x)}]_{\pi_{\gamma_0}(x)}\Big|_{(\alpha,\gamma)=(\alpha_0,\gamma_0)} \geq 0, \tag{12}$$

which ensure that (α_0,γ_0) is a relative minimum of (9).

A simple way to prove (9) is the use of the Gibbs inequality,

$$\sum_z P(z)\log\frac{Q(z)}{P(z)} \leq 0 \tag{13}$$

where $P(z)$ and $Q(z)$ are arbitrary functions that satisfy the relations $0 \leq P(z), Q(z) \leq 1$ and $\sum_z Q(z) = \sum_z P(z) = 1$. If we set $P(y) = \sum_x p_{\alpha_0}(y|x)\pi_{\gamma_0}(x)$ and $Q(y) = \sum p_\alpha(y|x)\pi_\gamma(x)$, it is easy to verify the requirement of the Gibbs inequality (13). Then it follows that, for any α and γ,

$$\left[\left[\log\frac{\sum_x p_\alpha(y|x)\pi_\gamma(x)}{\sum_x p_{\alpha_0}(y|x)\pi_{\gamma_0}(x)}\,\right]_{p_{\alpha_0}(y|x)}\right]_{\pi_{\gamma_0}(x)} \leq 0. \tag{14}$$

This proves (9) and its corollaries (10),(11),(12). We can also prove (10),(11),(12) through direct calculations similar to that for (5).

The history of Bayesian statistics is long and complicated. It originates from the Laplace's works on the "inverse probability" and has been an archetype of mathematical theories of uncertain objects. Despite different interpretations and objections to the use of prior distributions, it is an important language in a wide area of the sciences of information processing, say, time-series analysis [13, 16, 19], image restoration [20, 21, 22, 17, 23, 24, 18, 25], inference with neural networks [15, 26], and artificial intelligence. An earlier remark on the analogy between Bayesian statistics and statistical mechanics is found, for example, in Iba [27]. Sourlas [28] seems the first reference that discussed the relation between the coding theory and spin glasses. We also refer recent works of Bruce and Saad [29], Pryce and Bruce [30], Tanaka and Morita [31], Kabashima and Saad [32, 33], Dress et al. [34], Opper and Winther [35], which treat the relation between statistical mechanics and Bayesian statistics (or error-correcting codes).

3. The Nishimori Line

Now we discuss the relation between the results in the previous section and the Nishimori line of spin glasses. To see this in the simplest case of ±J Ising spin glass, we set the distributions as follows:

$$p_\alpha(y|x) = \frac{1}{Z_\alpha}\exp(-E_\alpha(x,y)), \tag{15}$$

$$-E_\alpha(x,y) = \alpha\sum_{(i,j)} y_{ij}x_ix_j, \tag{16}$$

$$Z_\alpha = \sum_y \exp(-E_\alpha(x,y)) = (\exp(\alpha)+\exp(-\alpha))^M, \tag{17}$$

and

$$\pi(x) = \frac{1}{2^N}\,(\text{the uniform distribution}), \tag{18}$$

where each of the component $x_i\,(i \in \{1..N\})$ of the parameter x takes the value of ±1. The component x_i is defined on the vertices i of a graph G, say a square lattice or a random network, of degree N. The data $y = \{y_{ij}\}$ is defined on the edges (i,j) of G and the summation $\sum_{(i,j)}$ runs over them. We denote the number of the edges of G as M, which is also the number of the data.

This probability $p_\alpha(y|x)$ corresponds to a binary symmetric channel where a set $\{y^{in}_{ij}\}\,((i,j) \in G)$ of the pair product $y^{in}_{ij} = x_ix_j$ of the inputs is sent as an error-correcting code [28, 10, 8, 32]. Here, "binary symmetric" means that the output of the channel y_{ij} is given by the formula

$$\begin{aligned} y_{ij} &= +y^{in}_{ij} \text{ with probability } 1-q \\ y_{ij} &= -y^{in}_{ij} \text{ with probability } q. \end{aligned} \tag{19}$$

If we assume that the data y_{ij} is generated by $p_{\alpha_0}(y|x)$ and $\pi(x)$ defined by (15) and (18), the noise q of the channel is related to the hyperparameter α_0 by

$$q = \frac{\exp(-\alpha_0)}{\exp(\alpha_0)+\exp(-\alpha_0)}\,. \tag{20}$$

Although "pair product code" $y^{in}_{ij} = x_ix_j$ defined in the above looks rather artificial one, recent works [32, 33] on error-correcting codes suggest its generalization might have practical importance §.

The posterior distribution of the model with data $\{y_{ij}\}$ and hyperparameter α is

$$p_\alpha(x|y) = \frac{1}{Z_{pos}}\exp(-E_\alpha(x,y)), \tag{21}$$

§ Another interesting interpretation of the probability $p_\alpha(y|x)$ is given by a problem [36] that arises in the analysis of social network data [37]. With this interpretation, the index i indicate a person and the binary variable $y_{ij} \in \{\pm1\}\,(y_{ij} = y_{ji})$ indicates a social relation between persons i and j, say, whether they have an acquaintance or not. Each person is assumed to subject to one of the social groups A and B, and the problem is to infer the group structure from the data $\{y_{ij}\}$. We set the indicator $x_i = 1$ when $i \in A$ and $x_i = -1$ when $i \in B$ and assume the following property:

> If a pair of the persons i and j is in the same social group, $y_{ij} = 1$ with a probability q and -1 with a probability $1-q$. Else if they are in different groups, $y_{ij} = 1$ with probability q' and -1 with a probability $1-q'$.

Then we get the probability $p_\alpha(y|x)$ in the text as a special case where $q' = 1-q$.

$$Z_{pos} = \sum_x \exp(-E_\alpha(x, y)). \tag{22}$$

This is the Gibbs distribution of a random bond Ising model with coupling constants $\{y_{ij}\}$ defined on the graph G. The derivatives of the function F defined by (8) are

$$\frac{\partial}{\partial\alpha} F(\alpha, \gamma) = -\frac{1}{\alpha}\langle E_\alpha(x', y)\rangle_{p_\alpha(x'|y)} - M \cdot \tanh\alpha \tag{23}$$

where $\langle\ \rangle_{p_\alpha(x|y)}$ indicates the canonical average with the energy $E(x, y)$ (Here and hereafter, we assume that we are working at the unit temperature $T = 1$ and α is treated as a (hyper)parameter of the model but not the temperature.). The term $M \cdot \tanh\alpha$ comes from the derivative of the logarithm of the normalization factor $(\exp(h) + \exp(-h))^M$ of the probability $p_\alpha(y|x)$.

In general, a misspeficification of hyperparameter α in (21) and (23) is possible. If we assume that we know the "true" value α_0 used in the generation of the data $\{y_{ij}\}$ and set $\alpha = \alpha_0$, or, equivalently,

$$\frac{\exp(-\alpha)}{\exp(\alpha) + \exp(-\alpha)} = q\,, \tag{24}$$

in the formulas, we have an identity on the energy

$$-[\langle E_\alpha(x', y)\rangle_{p_\alpha(x'|y)}]_{p_\alpha(y|x)}]_{\pi(x)} = \alpha M \cdot \tanh\alpha. \tag{25}$$

from the identity (10).

So far, the average over the bond randomness

$$[\ [\cdots]_{p_\alpha(y|x)}\ ]_{\pi(x)} \tag{26}$$

has a rather complicated structure. There are two steps in the generation of the quenched random variables $\{y_{ij}\}$, which are described by $\pi(x)$ and $p_\alpha(y|x)$ respectively. In this particular case, we can simplify it using *gauge invariance* of the problem. The gauge transformation group of this model is defined by the family of transformations,

$$V_z : \{y_{ij}\} \to \{z_i \cdot y_{ij} \cdot z_j\} \tag{27}$$

$$U_z : \{x_i\} \to \{x_i \cdot z_i\} \tag{28}$$

parameterized by $z = \{z_i\}, z_i \in \{\pm 1\}$. This set of transformations consist of one-to-one onto-mappings (permutations) of their domain and satisfy a transitive property, i.e., there exists z with which $U_z(x) = x'$ for any pair of x and x' in the domain. It is easy to show the following relations:

$$p_\alpha(U_z(x)|V_z(y)) = p_\alpha(x|y) \tag{29}$$

$$\pi(U_z(x)) = \pi(x) \tag{30}$$

$$p_\alpha(V_z(y)|U_z(x)) = p_\alpha(y|x), \tag{31}$$

$$E_\alpha(U_z(x), V_z(y)) = E_\alpha(x, y). \tag{32}$$

They are an example of the gauge invariance (or gauge covariance) in a random spin system ([6, 38, 1, 2], see [5] for a comprehensive treatment with applications to the Nishimori line.). By these formulas, we can show that the left hand side of the expression (25) is written as a simpler average

$$
\begin{aligned}
&[[\langle E_\alpha(x', y)\rangle_{p_\alpha(x'|y)}]_{p_\alpha(y|x)}]_{\pi(x)} \\
&= \sum_x \sum_y \sum_{x'} E_\alpha(x', y) \cdot p_\alpha(x'|y) \cdot p_\alpha(y|x) \cdot \pi(x) \\
&= \sum_x \sum_y \sum_{x'} E_\alpha(U_z(x'), V_z(y)) \cdot p_\alpha(U_z(x')|V_z(y)) \cdot p_\alpha(V_z(y)|U_z(x)) \cdot \pi(x) \\
&= \sum_y \sum_{x'} E_\alpha(x', y) \cdot p_\alpha(x'|y) \cdot p_\alpha(y|x^*) \\
&= [\langle E_\alpha(x', y)\rangle_{p_\alpha(x'|y)}]_{p_\alpha(y|x^*)} \qquad (33)
\end{aligned}
$$

where x^* is a ferromagnetic state $\{x_i^*\}$ ($\forall i \, x_i^* = 1$) and z is a function of x that satisfy $x^* = U_z(x)$. The existence of such z is secured by the transitive property and the change of the dummy index, say, from $V_z(y)$ to y in the summation $\sum_y$, is justified by the one-to-one onto property of the gauge transformations (27),(28) respectively.

Here and hereafter, we denote the average over the distribution

$$p_\alpha(y|x^*) = \frac{\exp(\alpha \sum_{(i,j)} y_{ij})}{Z_\alpha} \qquad (34)$$

as $[\ ]_q$, where the relation between q and α is defined in (24). It is easy to see a component y_{ij} of sample from the distribution (34) is mutually independent samples from the distribution

$$\Pr(y_{ij}) = q \cdot \delta(y_{ij} - 1) + (1 - q) \cdot \delta(y_{ij} + 1). \qquad (35)$$

By using (33) and these notations, the formula (25) is reduced to the identity

$$-[\langle E(x, y)\rangle_{p_\alpha(x|y)}]_q = \alpha M \cdot \tanh \alpha \qquad (36)$$

on the average of the energy of ±J spin glass with the coupling constants $\{y_{ij}\}$ from the distribution (35). It is nothing but a result reported in the first paper [1] on the Nishimori line.

The relation (24) between (hyper)parameter α in the canonical average and the noise level q in the quenched average is essential and known as the definition of the *Nishimori line* of the model. In our derivation, it arises from the condition $\alpha = \alpha_0$ in the formulas (9) and (10). This means that the model $p_\alpha(y|x)$ assumed in the analysis of the data coincides with the "true" probability $p_{\alpha_0}(y|x)$ used in the generation of the data. In terms of the coding theory, it corresponds to the situation where the decoder knows exactly the property of the channel, the coding, and the relative frequencies of the possible messages. It is rather surprising that the notion of the Nishimori line ‖, which is introduced without any background of statistical information processing, has such a clear interpretation from our point of view.

A similar argument with the substitution of the second derivative $\frac{\partial^2}{\partial^2 \alpha} F(\alpha, \gamma)$ of F into the expression (12) leads to an inequality

$$[[\,\langle E_\alpha^2(x', y)\rangle_{p_\alpha(x'|y)} - \langle E_\alpha(x', y)\rangle_{p_\alpha(x'|y)}^2\,]_{p_\alpha(y|x)}]_{\pi(x)} \le \frac{\alpha^2 M}{\cosh^2 \alpha}. \qquad (37)$$

‖ In general cases, where more than one (components of) hyperparameters are contained in the model, it is actually "Nishimori hypersurface". The term *Nishimori temperature* is also used by statistical physicists. It seems, however, not adequate terminology in the context of information processing, because the notion of temperature has no specific meaning in the problems in the statistics and the coding theory.

With gauge invariance of the model, we can derive an inequality

$$[\langle E_\alpha(x,y)^2\rangle_{p_\alpha(x|y)} - \langle E_\alpha(x,y)\rangle^2_{p_\alpha(x|y)}]_q \leq \frac{\alpha^2 M}{\cosh^2\alpha} \tag{38}$$

from (37). This is an inequality on the fluctuation of the energy (the specific heat) on the Nishimori line, which is also discussed in [1]. Some of the other relations that hold on the Nishimori line of the model is derived from the identity (7) and the gauge invariance of the model. For example, the distribution of the internal fields at the vertex i_0 [3] is reproduced, when we set $C(x,y) = \sum_j y_{i_0 j} x_j$, where j runs over the set of vertices neighboring to i_0, i.e., $(i_0, j) \in G$. The expression of the gauge invariant correlation function [1] is also derived, when we set $C(x,y) = x_k \cdot (\Pi_{(i,j)\in\Gamma}\ y_{ij}) \cdot x_l$, where Γ denotes a path that connects the vertices k and l.

Here we discuss a statistical model defined by (15) and (18), which leads to the Nishimori line of the ±J spin glass model. Our argument is, however, general and can be applied to the Nishimori line of other models, say, spin glasses with a Gaussian distribution of the coupling [2], models with multiple spin interactions [28, 8, 5, 34, 32, 33], and the gauge glasses [5, 9]. For each model, we can consider the corresponding statistical model (or a noisy channel) and derive the properties of the Nishimori line from the relations (10), (12), (7) with additional arguments on the gauge invariance. For example, we consider the following problem of statistical inference:

> There are a collection of the phase variable $\{x_i\}$ $(x_i \in [0, 2\pi)\ )$ on the vertices of a graph G, say, a two-dimensional lattice. The numbers of the vertices and edges in G are N and M, respectively. We assume that the difference $y^{in}_{ij} = x_i - x_j$ of the parameters x_i and x_j is observed for each edge (i,j) in G with a probabilistic error η_{ij}. The problem is to infer the original values of $\{x_i\}$ from the data $\{y_{ij}\}$ with the assumption that $y_{ij} = y^{in}_{ij} + \eta_{ij}$.

Such a problem might have practical importance in the analysis of the data from optical measurements where differences of the phase between neighboring points are observed with noise. If we assume that the magnitudes of the noise $\{\eta_{ij}\}$ are mutually independent and obey a von Mises distribution, a correspondence of a Gaussian distribution on a circle, the probability is

$$p_\alpha(y|x) = \frac{1}{Z_\alpha}\exp(-E_\alpha(x,y)), \tag{39}$$

$$-E_\alpha(x,y) = \alpha\sum_{(i,j)}\cos(x_i - x_j - y_{ij}), \tag{40}$$

$$Z_\alpha = \int \Pi_i\, dx_i\ \ p_\alpha(y|x) = (2\pi I_0(\alpha))^M, \tag{41}$$

where I_0 indicates the modified Bessel function. We also assume the uniform prior

$$\pi(x)dx = (\frac{1}{2\pi})^N \cdot \Pi_i\, dx_i, \qquad x_i \in [0, 2\pi). \tag{42}$$

From this setting, we can derive the results on the Nishimori line of the *gauge glass* [5, 9] with a method that is similar to that for the ±J spin glass.

4. What is New ?

Now, careful readers may ask *what is really new* in our approach. Once the expression

$$\sum_x \sum_y \left(\frac{\sum_{x'} (\sum_{(i,j)} y_{ij} x'_i x'_j) \cdot \exp(\alpha \sum_{(i,j)} y_{ij} x'_i x'_j)}{\sum_{x'} \exp(\alpha \sum_{(i,j)} y_{ij} x'_i x'_j)} \cdot \frac{\exp(\alpha \sum_{(i,j)} y_{ij} x_i x_j)}{Z_\alpha} \right) \qquad (43)$$

of the left hand side of (36) is derived by the gauge invariance from that of (25), it is not difficult to show the relation (36) by direct inspection of the expression. If we combined these steps of the proof in this order, it is nothing but a conventional proof [2] of the property of the Nishimori line. The same is true for the derivation with the identity (7). In this sense, our argument is not a re-derivation of the Nishimori line but a *reformulation* or *re-interpretation* of the original derivation.

There are, however, two major advantages of our approach. First, our interpretation elucidates the meaning of the Nishimori line. It is a line on which we make inference (or decoding) using the "true" probability structure that generates the data (or codes). This coincidence of the encoding and decoding scheme gives drastic simplifications of the averages of various kind of physical quantities. This interpretation also explains why there are two different variables x and x' with the same type in the expression (43). The variable x corresponds to the original value of the parameter (the input message) and x' represents an inference on it. The expression (43), in itself, has clear meaning as the relation (25) between averages. On the other hand, in the conventional derivation [2] of the Nishimori line, the insertion of the variable x to the left hand side of (36) looks a rather artificial procedure and the expression (43), which is defined in the configuration space enlarged by a gauge transformation, seems to have no definite meaning. This lacks of the interpretation is a reason why the derivation of the Nishimori line looks somewhat mysterious, even though the manipulation of the formula required in the proof is quite simple and elegant. We believe that our interpretation will contribute to make this point clear. Another interesting point of our argument is that it gives a novel interpretation to the identity (36) of the energy and the inequality (38). They are essentially necessary conditions that the average of the marginal likelihood takes the maximum value on the Nishimori line.

The second advantage of our approach is that it suggests the existence of the correspondence of the Nishimori line in the models *without gauge invariance.* The relations (10),(12), and (7), which are used in the derivation of the Nishimori line, are obtained without the gauge invariance of the models. Thus, we can prove the identity of the energy, inequality of the specific heat, and the expression of the distribution of internal fields etc., which hold on the "Nishimori line" of the models without gauge invariance. For example, we consider a Bayesian model

$$p_\alpha(y|x) = \frac{1}{Z_\alpha} \exp(-E_\alpha(x,y)), \qquad (44)$$

$$-E_\alpha(x,y) = \alpha \sum_i y_i x_i, \qquad (45)$$

$$Z_\alpha = \sum_y \exp(-E_\alpha(x,y)) = (\exp(\alpha) + \exp(-\alpha))^N, \qquad (46)$$

and

$$\pi_\gamma(x) = \frac{1}{Z_\pi} \exp(-E_\gamma(x,y)), \qquad (47)$$

$$-E_\gamma(x) = \gamma \sum_{(i,j)} x_i x_j, \tag{48}$$

$$Z_\gamma = \sum_x \exp(-E_\gamma(x)). \tag{49}$$

In this case, we assume that the unknown parameters $\{x_i\}$ and the data $\{y_i\}$ are defined on the vertices of a graph G with the degree N. When G is a two or three dimensional lattice, this model corresponds to an image restoration problem with a prior knowledge on images that is well described by the Ising prior (47)(For image restoration with Ising and Potts priors, see the references [20, 21, 22, 24, 25, 30, 31].). The posterior distribution of the model is

$$p_{\alpha\gamma}(x|y) = \frac{1}{Z_{\alpha\gamma}} \exp(-E_{\alpha\gamma}(x,y)), \tag{50}$$

$$-E_{\alpha\gamma}(x,y) = \alpha \sum_i y_i x_i + \gamma \sum_{(i,j)} x_i x_j, \tag{51}$$

$$Z_{\alpha\gamma} = \sum_x \exp(-E_{\alpha\gamma}(x,y)). \tag{52}$$

This is the Gibbs distribution of an Ising model with an inhomogeneous external field $\{h \cdot y_i\}$.

Let us consider the cases where the data generation mechanism is exactly described by the probabilities (44) with $\alpha = \alpha_0$ and (47) with $\gamma = \gamma_0$, i.e., the pattern of the random field $\{y_i\}$ is given by the following process : (i) Generate a sample pattern $\{y_i^{in}\}$ from the Gibbs distribution of the Ising model (47) with the coupling constant γ_0. (ii) Flip each component $\{y_i^{in}\}$ with the probability

$$q = \frac{\exp(-\alpha_0)}{\exp(\alpha_0) + \exp(-\alpha_0)} \tag{53}$$

(a binary symmetric channel). Then, we can interpret the posterior (50) as a Gibbs distribution of a Random Field Ising Model (RFIM). Note that external fields on the sites of this model are not mutually independent random variables, but correlated with a way specified by (i) and (ii). The "Nishimori line" of this model is defined as a surface where the parameters (α_0, γ_0) in the definition of the quenched randomness coincide with (α, γ) in the canonical average. Equivalently,

$$\frac{\exp(-\alpha)}{\exp(\alpha) + \exp(-\alpha)} = q, \tag{54}$$

$$\gamma = \gamma_0, \tag{55}$$

where α and γ are the parameters in (51), and q and γ_0 are the parameters in (i) and (ii). From (10)(11) and (5) with $A(x) = x_i x_j$, we can prove identities that holds on the Nishimori line of this model,

$$[[\,\langle \sum_i y_i x_i \rangle_{p_{\alpha\gamma}(x|y)}\,]_{p_\alpha(y|x)}]_{\pi_\gamma(x)} = N \cdot \tanh\alpha, \tag{56}$$

$$[[\,\langle \sum_{(i,j)} x_i x_j \rangle_{p_{\alpha\gamma}(x|y)}\,]_{p_\alpha(y|x)}]_{\pi_\gamma(x)} = \langle \sum_{(i,j)} x_i x_j \rangle_\gamma^{pure}, \tag{57}$$

$$[[\,\langle x_i x_j \rangle_{p_{\alpha\gamma}(x|y)}\,]_{p_\alpha(y|x)}]_{\pi_\gamma(x)} = \langle x_i x_j \rangle_\gamma^{pure}. \tag{58}$$

Here and hereafter, the average $\langle \cdots \rangle_\gamma^{pure}$ is the canonical average with the "pure" Ising model with homogeneous couplings of the strength γ on the same graph G.

The expression (58) shows that the quenched average of the correlation of spins in the RFIM with a correlated random field is just the same as that of the corresponding pure Ising model. Furthermore, an identity between the order parameters is obtained if we consider a set of systems of a fixed boundary condition with which $x_i = 1$ for the all spins at the boundary [2]. That is, the random field $\{y_i\}$ is assumed to be generated by the process (i) and (ii) in which the values of the boundary spins are kept to 1, and the thermal averages $\langle x_i x_j \rangle_{p_{\alpha\gamma}(x|y)}$ and $\langle x_i x_j \rangle_\gamma^{pure}$ are also taken with this boundary condition. Assuming that the site j belongs to the boundary and the site i is located far from the boundary, the relation

$$[[\, \langle x_i \rangle_{p_{\alpha\gamma}(x|y)}\,]_{p_\alpha(y|x)}]_{\pi_\gamma(x)} = m_\gamma^{pure} \tag{59}$$

is derived from (58), where m_γ^{pure} is the bulk magnetization per spin of the corresponding pure system.

Although these results are dependent on the special features of the model, similar arguments are applicable in other models without gauge invariance and leads to identities and inequalities on the Nishimori line of the model. An example is provided by the posterior distribution corresponding to a binary asymmetric channel, which is already discussed in Sourlas [10] in the context of optimal decoding. It is related to models with a special type of site/bond randomness.

There are, however, some intrinsic limitations on the utility of the notion of the Nishimori line without gauge invariance. First, we cannot simplify the definition of the quenched average $[\, [\cdots]_{p_\alpha(y|x)}\,]_{\pi_\gamma(x)}$ on the Nishimori line without gauge invariance. Then, the results usually contain a complicated quenched average, which often lacks a clear correspondence to that in physical systems. The two-stage process (i) and (ii) of the generation of quenched randomness in the RFIM just described above is a typical example of this. Another important remark is that not all of the arguments on the Nishimori line with gauge invariance is applicable to the models without gauge invariance. For example, the identity

$$[\, \langle x_i' \rangle_{p_\alpha(x_i'|y)} \cdot \langle x_i' \rangle_{p_{\alpha'}(x'|y)}\,]_q = [\, \langle x_i' \rangle_{p_{\alpha'}(x'|y)}\,]_q \tag{60}$$

valid for the ±J spin glass model [2] has no correspondence in a model without gauge invariance. Here α and q is related by the condition (24) of the Nishimori line and α' takes an arbitrary value. The identity (60) is important because the upper bound

$$|[\, \langle x_i' \rangle_{p_{\alpha'}(x'|y)}\,]_q| \leq [|\langle x_i \rangle|_{p_\alpha(x|y)}]_q \tag{61}$$

of the order parameter is derived from it. If we substitute $C(x,y) = x_i \cdot \langle x_i' \rangle_{p_{\alpha'}(x_i'|y)}$ in (7), we can prove the relation

$$[[\, \langle x_i' \rangle_{p_\alpha(x_i'|y)} \cdot \langle x_i' \rangle_{p_{\alpha'}(x'|y)}\,]_{p_\alpha(y|x)}]_{\pi_\gamma(x)} = [[\, x_i \cdot \langle x_i' \rangle_{p_{\alpha'}(x'|y)}\,]_{p_\alpha(y|x)}]_{\pi_\gamma(x)}, \tag{62}$$

which apparently corresponds to (60). However, further simplification of the right hand side is not possible without gauge invariance. Unfortunately, the expression (62) gives little information on the shape of the boundaries in phase diagram and seems not so useful as (60).

5. Finite-temperature Decoding

The notion of the optimality of "finite temperature decoding" is introduced to the community of statistical physicists by Ruján [7] and discussed by Nishimori [8, 9] and

Sourlas [10]. Recently, it again draws an attention of the researchers of this area, because the development on the statistical mechanics of error-correcting codes [32] enables the quantitative argument of the problem with analytical methods.

Roughly speaking, "the optimality of finite temperature decoding" means that the estimator that maximize the posterior probability (Maximum A Posteriori estimator) is not always the best estimator. The best estimator is dependent on the purpose of inference (or decoding) and often defined with averages over the posterior distribution. If we call MAP estimator, which is defined as a "ground state" of the corresponding physical system, the "zero-temperature decoder", it is natural to call an estimator defined with the posterior averages a "finite temperature decoder" or "$T = 1$ decoder" [10].

This fact, however, had been well known in the study of the statistics and pattern recognition. For example, Marroquin [39](see also [24, 25]) discussed an estimator ("MPM estimator") in image restoration problems, which is just the same as the one proposed by Ruján [7]. Moreover, it is not the first work that uses the estimator in this field ¶. General arguments on the optimality of the estimator in the Bayesian framework is already found in the textbooks [40, 41, 42, 12] of statistics. The branch of statistics that discuss optimal decisions with uncertain information is known as *statistical decision theory.*

Here, we will briefly discuss the basic results on optimal estimators. Our treatment is not very different from the arguments in Sourlas [10] and those in the textbooks of statistics. It is, however, useful to give a coherent derivation with the notations in the earlier sections, because no comprehensive treatment on this subject seems available in the literature of physics.

To give a formal definition of optimal estimators, we introduce the notion of a *loss function* $L(x, \hat{x})$ that gives a measure of distance [+] between the original parameter x and an estimate $\hat{x}$ of x. Then we define an optimal estimator $\hat{x}(y)$ for a loss function L as a function of y that minimize the expected loss

$$[[L(x, \hat{x}(y))]_{p(y|x)}]_{\pi(x)}. \tag{63}$$

Here and hereafter we assume that we know exactly about the data generation process and omit the subscripts that indicate hyper parameters α, γ in the expressions, say, $p(y|x)$, $\pi(x)$ and $\langle\ \rangle$ (i.e., we set the values of the hyperparameters to their "true" values.). Note that the optimality of an estimator defined here is a very strong notion. It means that $\hat{x}(y)$ has better or equal average performance against any function of the data y, provided that the data generation scheme (or the set of the channel and the frequencies of the messages) is exactly described by the given probability $p(y|x)$ and $\pi(x)$. It is not restricted to the optimality in a series of the estimators defined with different hyperparameters or temperatures.

A basic result on optimal estimators is as follows:

> **Lemma**
> An optimal estimator $\hat{x}(y)$ for a loss function L is an estimator that minimize the posterior average $\langle L(x, \hat{x}(y))\rangle_{p(x|y)}$ for each y.

This result is a rather obvious one with the principle that the posterior distribution is the source of the all information from the data. If we use the identity (7), the formal

¶ See, for example, the references [23] and Sec 2.4 of [21]. A recent work on the optimal estimator in image restoration is found in Rue [43].

[+] It is *not* necessary to satisfy the axiom of the distance.

Table 1. Loss Functions and Corresponding Optimal Estimators.
If there are no special comments in the table, a component of parameters x_i and its estimate $\hat{x}_i$ are assumed to take their values in a subset of $\mathbf{R}$ and $\mathbf{R}$ respectably. The symbol $\langle\ \rangle$ indicates the average over the posterior distribution. An expression $\arg\max_z f(z)$ indicates a value of z that maximizes $f(z)$ and Kronceker delta $\delta_{w,z}$ is defined as usual, i.e., $\delta_{w,z} = 1$ if $w = z$, else $\delta_{w,z} = 0$.

loss function (L)	optimal estimator ($\hat{x}$)	comments
$\sum_i (x_i - \hat{x}_i)^2$	$\hat{x}_i = \langle x_i \rangle$	
$\sum_i \lvert x_i - \hat{x}_i \rvert$	$\hat{x}_i =$ median of $p_i(x_i)$ [a]	
$1 - \prod_i \delta_{x_i,\hat{x}_i}$	$\{\hat{x}_i\} = \arg\max_x p(x\lvert y)$ [b]	x_i : a discrete variable
$\sum_i (1 - \delta_{x_i,\hat{x}_i})$	$\hat{x}_i = \arg\max_{x_i} p_i(x_i)$ [a]	x_i : a discrete variable
$-\sum_i x_i \hat{x}_i$	$\hat{x}_i = \frac{\langle x_i \rangle}{\lvert\langle x_i \rangle\rvert}$	$\hat{x}_i \in \pm 1$
$-\sum_i \{x_i \log \hat{x}_i + (1 - x_i)\log(1 - \hat{x}_i)\}$	$\hat{x}_i = \langle x_i \rangle$	$0 < x_i, \hat{x}_i < 1$

[a] Here $p_i(x_i)$ indicates the marginal distribution $p_i(x_i) = \sum_{\{x_j\}\, j\neq i} p(x|y)$ of x_i, where $\sum_{\{x_j\}\, j\neq i}$ means the summation over x with a fixed value of the ith component x_i.
[b] It is often called "MAP (Maximum A Posteriori) estimator".

derivation of the lemma is easy. When we set $C(x, y) = L(x, \hat{x}(y))$, the expression (7) gives

$$[[\langle L(x', \hat{x}(y))\rangle_{p(x'|y)}]_{p(y|x)}]_{\pi(x)} = [[L(x, \hat{x}(y))]_{p(y|x)}]_{\pi(x)}. \qquad (64)$$

Here we note three observations on the formulas (64): (a) The estimator $\hat{x}(y)$ is an arbitrary function of y and we can freely attribute its value at each y. (b) The average $[\,[\cdots]_{p(y|x)}\,]_{\pi(x)}$ in the left hand side of (64) is an average over y with non-negative weights. (c) The function $L(x', \hat{x}(y))$ does not explicitly contains the original parameter x. By using (a),(b) and (c), we can see that the minimizer of the left side hand of (64) is the minimizer of the posterior average $\langle L(x, \hat{x}(y))\rangle_{p(x|y)}$ for each y. Thus, the lemma is proved.

For example, consider the case where the distance L between the binary sequence $x = \{x_i\}$ and $\hat{x} = \{\hat{x}_i\}\,(x_i, \hat{x}_i \in \{\pm 1\})$ is measured by the overlap $\sum_i \hat{x}_i x_i$ of the pattern, i.e., $L(x, \hat{x}) = -\sum_i \hat{x}_i x_i$. With this loss function,

$$\langle L(x, \hat{x})\rangle_{p(x|y)} = -\sum_i \hat{x}_i(y)\langle x_i\rangle_{p(x|y)}. \qquad (65)$$

Then, the optimal estimator $\hat{x}_i(y) \in \{\pm 1\}$, which minimize the right hand side of (65), is given by

$$\hat{x}_i(y) = \frac{\langle x_i\rangle_{p(x|y)}}{|\langle x_i\rangle_{p(x|y)}|}. \qquad (66)$$

This expression coincides with the result in [7, 8, 10, 39, 24, 25]. Examples of loss functions and the corresponding optimal estimators are shown in the Table 1. By using the lemma, we can easily derive them.

So far, our discussion in this section does not depend on the notion of the gauge invariance. Correspondences between loss functions and optimal estimators shown in the Table 1 are independent of the existence of gauge invariance of the model. With gauge invariance, we can prove an additional result. Let us assume that the model is gauge invariant and the following properties of the loss function L and the estimator $\hat{x}(y)$

$$L(\,U_z(x), U_z(\hat{x})\,) = L(x, \hat{x}) \tag{67}$$

$$U_z(\,\hat{x}(y)\,) = \hat{x}(\,V_z(y)\,) \tag{68}$$

are satisfied for all z (The mappings V_z and U_z are defined in the section 3.). Then, we can show that the expected loss $[L(x, \hat{x}(y))]_{p(y|x)}$ with any fixed x is independent of the value of x. The proof ([12] p.396, [42] p.168) is as follows:

$$\begin{aligned}
[L(x, \hat{x}(y))]_{p(y|x)} &= \sum_y L(x, \hat{x}(y)) \cdot p(y|x) \\
&= \sum_y L(x^*, U_z(\hat{x}(y))) \cdot p(V_z(y)|x^*) \\
&= \sum_y L(x^*, \hat{x}(V_z(y))) \cdot p(V_z(y)|x^*) \\
&= \sum_y L(x^*, \hat{x}(y)) \cdot p(y|x^*) \\
&= [L(x^*, \hat{x}(y))]_{p(y|x^*)}
\end{aligned} \tag{69}$$

where x^* is an arbitrary chosen "standard" configuration, say a ferromagnetic state, and z is chosen to satisfy the relation $x^* = U_z(x)$. The result (69) means that the estimator performs equally well for any value of the original parameter x. In terms of statistics [42], an estimator that is optimal within the class of the estimators with such uniformity is called a minimum risk equivariant estimator (MRE) [*]. The case discussed in Ruján [7] and Nishimori [8] corresponds to a special example of MRE.

In fact, we can remove the assumption (68) on the estimator, if the estimator is optimal and the optimal estimator is known to be unique. This means that if the loss function is gauge invariant, the corresponding optimal estimator is automatically gauge covariant and satisfies (68) [♯] . The proof is easy, if we note that the estimator defined by

$$\hat{x}_z(y) = U_z^{-1}(\,\hat{x}(V_z(y))\,) \tag{70}$$

is an estimator of the equal performance to the original estimator $\hat{x}(y)$, i.e.,

$$[[L(x, \hat{x}_z(y))]_{p(y|x)}]_{\pi(x)} = [[L(x, \hat{x}(y))]_{p(y|x)}]_{\pi(x)}. \tag{71}$$

The relation (71) is confirmed by the calculation similar to that in the proof of (69) under the assumption of (67) and the gauge covariance of $p(y|x)$ and $\pi(x)$. Thus, with the assumption of the uniqueness of the optimal estimator, $\hat{x}_z(y)$ should be coincides with $x(y)$ for any value of z. It proves the relation (68).

[*] The term "invariant" is also used. The author prefers "covariant", but does not know whether it has been used by statisticians. Here we restrict ourselves within the special form of U_z and V_z induced by the gauge transformation group of Ising spin glass. See the reference [12, 42] for definitions and results with an arbitrary group of transitive transformations.

[♯] It is not true without an additional assumption on the uniqueness. A counter example is given by a binary symmetric channel with extreme noise $q = 1/2$, which transmits no information. For this example, any estimator is "optimal" for $L(x, \hat{x}) = -\sum_i \hat{x}_i x_i$. Some of them, say an estimator that returns a constant as estimates, are evidently not a MRE.

Acknowledgments

The author would like to thank Y. Kabashima and D. Saad for fruitful discussions and kind advices.

References

[1] Nishimori H 1980 Exact results and critical properties of the Ising model with competing interactions, *J. Phys. C: Solid State Phys.* **13** 4071–4076.

[2] Nishimori H 1981 Internal energy, specific heat and correlation function of the bond-random Ising model, *Progress of Theoretical physics* **66** 1169–1181.

[3] Nishimori H 1986 Exact results on the Ising spin glass in finite dimensions, *Progress of Theoretical physics* **76** 305–306.

[4] Nishimori H 1992 Boundary between the Ferromagnetic and spin glass phases, *J. Phys. Soc. Japan* **61** 1011–1012.

[5] Ozeki Y and Nishimori H 1993 Phase diagram of gauge glasses, *J. Phys. A: Math. Gen.* **26** 3399.

[6] Toulouse G 1977 Theory of the frustration effect in spin glasses: I , *Comm.Phys.* **2** 115-119 reprinted in *Spin Glasses and Beyond* 1987 eds. Mezard M, Parisi G and Virasoro M A (Singapore: World Scientific)

[7] Ruján P 1993 Finite temperature error-correction codes, *Phys. Rev. Lett.* **70** 2968–2971.

[8] Nishimori H 1993 Optimum decoding temperature for error-correcting codes, *J. Phys. Soc. Japan* **62** 2973–2975.

[9] Nishimori H 1993 Gauge glass, spin glass and coding theory – Exact results, *Proceedings of 2nd Taipei International Symposium on Statistical Physics* TITCMT-93-27.

[10] Sourlas N 1994 Spin-glasses,error-correcting codes and finite temperature decoding, *Europhysics letters* **25** 159–164.

[11] Good I J 1965 *The Estimation of Probabilities* (Cambridge Mass. : MIT Press).

[12] Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis* 2nd ed. (New York Berlin Heidelberg Tokyo: Springer-Verlag).

[13] Akaike H 1980 Likelihood and Bayes procedure in *Bayesian Statistics* Eds. Bernardo J M DeGroot M H Lindley D V and Smith A F M (Valencia: University press) p.143.

[14] MacKay D J C 1992 Bayesian Interpolation *Neural Computation* **4** 415–447.

[15] MacKay D J C 1992 A practical Bayesian framework for backprop networks, *Neural Computation* **4** 448–472.

[16] Kitagawa G and Gersch W 1996 *Smoothness Priors Analysis of Time Series* Lecture Notes in Statistics No.116 (New York: Springer-Verlag).

[17] Geman S and McClure D E 1987 Statistical methods for tomographic image reconstruction, *Proc. of the 46th Session of the ISI* Bulletin of the ISI **52** (Bulletin of the International Statistical Institute) p.5–21.

[18] Devijver P A and Dekesel M M 1987 Learning the parameters of a hidden Markov random field image model: A simple example, *Pattern Recognition Theory and Applications* NATO ASI series F30 eds. Devijver P A and Kittler J (Berlin: Springer-Verlag).

[19] West M and Harrison J 1997 *Baysian Forecasting and Dynamic Models* 2nd ed. (New York: Springer-Verlag).

[20] Besag J 1989 Towards Bayesian image analysis, *Journal of Applied Statistics* **16** 395–407.

[21] Besag J 1986 On the statistical analysis of dirty pictures, *The Journal of the Royal Statistical Society B* **48** 259–302 (with discussion).

[22] Geman S and Geman D 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

[23] Derin H, Elliott H, Cristi R and Geman D 1984 Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 707–720.

[24] Marroquin J, Mitter S and Poggio T 1987 Probabilistic solution of ill-posed problems in computational vision, *Journal of the American Statistical Association* **82** 76–89.

[25] Winkler G 1995 *Image analysis, Random fields and Dynamic Monte Carlo Methods, A Mathematical Introduction* (Berlin: Springer-Verlag) .

[26] Neal R M 1996 *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics, No.118, (New York: Springer).

[27] Iba Y 1989 Bayesian statistics and statistical mechanics, *Cooperative Dynamics in Complex Physical Systems*, ed. Takayama H (Berlin: Springer-Verlag) p.235–236

[28] Sourlas N 1989 Spin-glass models as error correcting codes, *Nature* **339** 693–695.

[29] Bruce A D and Saad D 1994 Statistical mechanics of hypothesis evaluation, *J. Phys. A: Math. Gen.* **27** 3355–3363.

[30] Pryce J M and Bruce A D 1995 Statistical mechanics of image restoration, *J. Phys. A: Math. Gen.* **28** 511–532.

[31] Tanaka K and Morita T 1995 Cluster variation method and image restoration problem, *Phys. Lett.* **A203** 122–128.

[32] Kabashima Y and Saad D 1997 Statistical mechanics of error-correcting codes, preprint.

[33] Kabashima Y and Saad D 1998 Belief propagation .vs. TAP for decoding corrupted messages, submitted to *Europhysics Letters*.

[34] Dress C, Amic E and Luck J M Zero-temperature error-correcting code for a binarysymmetric channel, *J. Phys. A: Math. Gen.* **28** 135–147.

[35] Opper M and Winther O (1996) A mean field approach to Bayes learning in feed-foward neural networks *Phys. Rev. Lett.* **76** 1964.

[36] Iba Y 1996 Markov chain Monte Carlo algorithm and their applications to statistics, *Proceedings of the Institute of Statistical Mathematics* **44** 49–84 (in Japanese with English Summary).

[37] Wang Y J and Wong G Y 1987 Stochastic blockmodels for directed graphs, *Journal of the American Statistical Association* **82** 8–19.

[38] Fradkin E, Huberman B A and Shenker S H 1978 Gauge symmetries in random magnetic systems, *Phys. Rev.* **B18** 4789–4814.

[39] Marroquin J 1985 Optimal Bayesian estimators for image segmentation and surface reconstruction, MIT A.I. Memo. No.839.

[40] Wonnacott T H and Wonnacott R W 1977 *Introductory Statistics* 3rd ed. (USA: John Wileys & Sons) ; see chapter 20 on the Bayesian decision theory.

[41] Matsubara N 1992 *Analysis of Scientific data : Basic Statistics 3* (in Japanese) ed. Statistics Section, Department of Social Science, College of Arts and Sciences (University of Tokyo, University of Tokyo Press) ; see chapter 9 on the Bayesian decision.

[42] Lehmann E L 1983 *Theory of Point Estimation* (California, USA: Wadsworth Inc. Reprinted in 1991).

[43] Rue H 1995 New loss functions in Bayesian imaging, *Journal of the American Statistical Association* **90** 900–908.